clave unica _____

Problema de Multi-Bandas (Multi-Armed Bandit): Teoría e Implementación

La tarea se entrega por discord antes del miercoles de la siguiente clase. Incluye llenar cuidadosamente en latex todos los snippets mencionados aqui, mas el codigo ya sea con link a colab o al repositorio. No olviden poner su clave unica arriva.

La proxima clase vamos a continuar con un ejercicio parecido, pero usando cadenas de markov. Vamos a modificar el bandit para que sea mas interesante ante cadenas de markov.

El lunes hay examen sobre estos ejercicios a papel y lapiz, para cada problema le ahre una pequenna modificacion y tendran que responder como lo resolverian.

1. Introducción a los Problemas de Multi-Bandas

1.1 Definición y Enunciado del Problema

El problema de Multi-Bandas (MAB, por sus siglas en inglés) es un problema clásico en teoría de la decisión y aprendizaje por refuerzo. Su nombre surge del escenario de un jugador que enfrenta múltiples máquinas tragamonedas (a veces llamadas "bandidos de un solo brazo"), cada una con diferentes probabilidades de recompensa desconocidas. El jugador debe decidir qué máquinas jugar, en qué orden y cuántas veces, para maximizar su recompensa total.

En este modelo:

- Existen K brazos (o acciones) diferentes.
- Cada brazo, cuando se jala, otorga una recompensa extraída de una distribución de probabilidad específica de ese brazo.
- Las distribuciones de recompensa son inicialmente desconocidas para el tomador de decisiones.
- El objetivo es maximizar la recompensa acumulada a lo largo de una serie de jugadas.

PROF

El problema captura la disyuntiva fundamental entre **exploración** (probar diferentes brazos para reunir información sobre sus distribuciones de recompensa) y **explotación** (elegir el brazo que actualmente parece ser el mejor).

1.2 Dilema de Exploración vs. Explotación

Este dilema está en el corazón del problema de multi-bandas:

- **Exploración**: Seleccionar brazos para aprender más sobre sus distribuciones de recompensa, potencialmente sacrificando recompensas inmediatas.
- **Explotación**: Seleccionar el brazo que actualmente parece ofrecer la mayor recompensa esperada en función de la información reunida hasta el momento.

Equilibrar estos dos aspectos es crucial. Demasiada exploración desperdicia recursos en brazos subóptimos. Demasiada explotación puede impedir descubrir un brazo mejor.

1.3 Formulación Matemática General

Formalicemos el problema estándar de bandas estocásticas:

- Sea K el número de brazos.
- Para cada brazo $i \in {1, 2, \ldots, K}$, existe una distribución de probabilidad desconocida $\mathcal{D}_i$ con media μ_i.
- En cada paso de tiempo $t \in {1, 2, \ldots, T}$:
    - El agente selecciona un brazo $a_t \in {1, 2, \ldots, K}$.
    - El agente recibe una recompensa $r_t \sim \mathcal{D}_{a_t}$.
- El objetivo es maximizar la recompensa acumulada $\sum_{t=1}^{T} r_t$.

Alternativamente, el problema puede enmarcarse en términos de minimizar **el arrepentimiento**. El arrepentimiento se define como la diferencia entre la recompensa obtenida al seleccionar siempre el brazo óptimo y la recompensa realmente obtenida por el agente:

$\text{Regret}(T) = T \cdot \max_{i} \mu_i - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$

2. Escenarios de Información en Nuestro Entorno de Bandas

En nuestro entorno de multi-bandas, exploramos tres escenarios de información distintos, cada uno proporcionando al agente diferentes niveles de conocimiento:

2.1 Escenario de Información Completa

En este escenario, el agente observa:

- El número de turno actual.
- El número total de turnos T.
- La probabilidad de recompensa para el brazo 1 (p1).
- El historial completo de acciones y recompensas pasadas.

PROF

Este es el escenario más informativo, ya que el agente conoce la probabilidad de uno de los brazos directamente y puede inferir la del otro con base en las recompensas observadas.

2.2 Escenario de Información Parcial

En este escenario, el agente observa:

- El número de turno actual.
- El número total de turnos T.
- La probabilidad de recompensa para el brazo 1 (p1).
- El historial de acciones y recompensas pasadas.

El agente conoce la probabilidad de un brazo pero debe aprender la del otro a través de la experimentación.

2.3 Escenario de Solo Recompensa

En este escenario, el agente observa:

- El número de turno actual.
- El historial de acciones y recompensas pasadas.

Este es el escenario más desafiante porque:

1. El agente no conoce la probabilidad de ninguno de los dos brazos.
2. El agente no conoce el número total de turnos T.

El agente debe aprender las probabilidades de ambos brazos mediante la experimentación y no puede optimizar su estrategia en función de la duración conocida del juego.

3. Entornos de Bandas en Nuestro Playground

Nuestro entorno implementa cuatro tipos diferentes de entornos de multi-bandas, cada uno con características distintas que afectan cómo cambian las probabilidades de los brazos a lo largo del tiempo.

3.1 Entorno de Banda Fija

Descripción

En el entorno de Banda Fija, cada brazo tiene una probabilidad constante de recompensa durante todo el juego. Estas probabilidades se asignan aleatoriamente al inicio de cada juego (uniforme entre 0.01 y 0.99) y permanecen sin cambios.

Formulación Matemática

- Dos brazos: $a \in {0, 1}$
- Probabilidades fijas: $p_1, p_2 \in [0.01, 0.99]$
- En el turno t, al seleccionar el brazo a:
    - Se recibe recompensa $r_t = 1$ con probabilidad p_{a+1}
    - Se recibe recompensa $r_t = 0$ con probabilidad $1 - p_{a+1}$

Decisión (T Fijo)

EJERCICIO

RESPUESTA

Definir el problema de decisión para la Banda Fija con horizonte de tiempo conocido T = 100. ¿Cuál es la función objetivo? ¿Cuáles son las restricciones? ¿Cuál es la política óptima?

PROF

Decisión (T Aleatorio)

EJERCICIO

RESPUESTA
Definir el problema de decisión para la Banda Fija con horizonte de tiempo desconocido T ~ Uniform(1, 300). ¿Cómo afecta el horizonte de tiempo aleatorio la estrategia óptima?

3.2 Entorno de Banda Periódica

Descripción

En el entorno de Banda Periódica, la probabilidad de recompensa de cada brazo cambia cada k turnos (por defecto, k=10). En cada punto de cambio, se asignan nuevas probabilidades aleatorias (uniforme entre 0.01 y 0.99) a ambos brazos.

PROF

Formulación Matemática

- Dos brazos: $a \in {0, 1}$
- En el turno t, las probabilidades son:
  - $p_1(t) = p_1^{\lfloor t/k \rfloor}$, donde $p_1^j \sim \text{Uniform}(0.01, 0.99)$
  - $p_2(t) = p_2^{\lfloor t/k \rfloor}$, donde $p_2^j \sim \text{Uniform}(0.01, 0.99)$
- El superíndice $j = \lfloor t/k \rfloor$ indica el número de "período".
- En cada punto de cambio (cuando t es divisible por k), se asignan nuevos valores aleatorios.

Decisión (T Fijo)

EJERCICIO

RESPUESTA
Definir el problema de decisión para la Banda Periódica con horizonte de tiempo conocido T = 100 y

período k = 10. ¿Cómo abordarías la búsqueda de una estrategia óptima? ¿Qué información adicional sería valiosa rastrear?

Decisión (T Aleatorio)

EJERCICIO

RESPUESTA
Definir el problema de decisión para la Banda Periódica con horizonte de tiempo desconocido T ~ Uniform(1, 300) y período k = 10. ¿Cómo interactúa la aleatoriedad en T con la naturaleza periódica del entorno?

PROF

3.3 Entorno de Banda Dinámica

Descripción

En el entorno de Banda Dinámica, las probabilidades de recompensa para ambos brazos cambian en cada turno. Cada turno se asignan probabilidades aleatorias completamente nuevas (uniforme entre 0.01 y 0.99) a ambos brazos.

Formulación Matemática

- Dos brazos: $a \in {0, 1}$
- En el turno t, las probabilidades son:
  - $p_1(t) \sim \text{Uniform}(0.01, 0.99)$
  - $p_2(t) \sim \text{Uniform}(0.01, 0.99)$
- Se generan nuevos valores aleatorios en cada turno.

Decisión (T Fijo)

EJERCICIO

RESPUESTA
Definir el problema de decisión para la Banda Dinámica con horizonte de tiempo conocido T = 100. ¿Hay una forma significativa de aprender de observaciones pasadas en este entorno? ¿Cuál sería la estrategia óptima?

Decisión (T Aleatorio)

EJERCICIO

RESPUESTA
Definir el problema de decisión para la Banda Dinámica con horizonte de tiempo desconocido T ~ Uniform(1, 300). ¿Cambia significativamente el enfoque óptimo en este entorno altamente dinámico si el horizonte de tiempo es desconocido?

PROF

3.4 Entorno de Banda Totalmente Aleatorio

Descripción

En el entorno de Banda Totalmente Aleatorio, las probabilidades de los brazos se inicializan de forma aleatoria y luego cambian aleatoriamente con una pequeña probabilidad (5%) en cada turno. Esto crea un entorno donde los cambios son impredecibles pero ocurren con menos frecuencia que en el entorno Dinámico.

Formulación Matemática

- Dos brazos: $a \in {0, 1}$
- Probabilidades iniciales: $p_1(0), p_2(0) \sim \text{Uniform}(0.01, 0.99)$
- En el turno $t > 0$, con probabilidad 0.05:
    - $p_1(t) \sim \text{Uniform}(0.01, 0.99)$
    - $p_2(t) \sim \text{Uniform}(0.01, 0.99)$
- De lo contrario (con probabilidad 0.95):
    - $p_1(t) = p_1(t-1)$
    - $p_2(t) = p_2(t-1)$

Decisión (T Fijo)

EJERCICIO

RESPUESTA

Definir el problema de decisión para la Banda Totalmente Aleatoria con horizonte de tiempo conocido T = 100. ¿Cómo equilibrarías la exploración y explotación sabiendo que las probabilidades de los brazos podrían cambiar repentinamente?

Decisión (T Aleatorio)

EJERCICIO

PROF

RESPUESTA

Definir el problema de decisión para la Banda Totalmente Aleatoria con horizonte de tiempo desconocido T ~ Uniform(1, 300). ¿Cómo interactúan las dos formas de aleatoriedad (en las probabilidades de los brazos y en el horizonte de tiempo)?

4. Implementación de Agentes

En nuestro entorno, implementarás tres tipos de agentes correspondientes a los tres escenarios de información descritos anteriormente. Esto es lo que cada agente debe manejar:

4.1 Agente de Información Completa

Entrada:

```
env_info = {
    'current_turn': int,        # Número de turno actual
    'total_turns': int,         # Número total de turnos en el juego
    'p1': float,                # Probabilidad de recompensa del brazo 1
    'history': {
        'actions': [int, ...],   # Acciones pasadas (0 para brazo 1, 1
para brazo 2)
        'rewards': [float, ...], # Recompensas pasadas
        'p1': [float, ...],      # Historial de probabilidades del brazo
1
        'p2': [float, ...]       # Historial de probabilidades del brazo
2 (solo para evaluación)
    }
}
```

Salida:

```
action = 0 or 1  # 0 para el brazo 1, 1 para el brazo 2
```

4.2 Agente de Información Parcial

PROF

Entrada:

```
env_info = {
    'current_turn': int,        # Número de turno actual
    'total_turns': int,         # Número total de turnos en el juego
    'p1': float,                # Probabilidad de recompensa del brazo 1
    'history': {
        'actions': [int, ...],   # Acciones pasadas (0 para brazo 1, 1
para brazo 2)
        'rewards': [float, ...]  # Recompensas pasadas
    }
}
```

Salida:

```
action = 0 or 1  # 0 para el brazo 1, 1 para el brazo 2
```

4.3 Agente de Solo Recompensa

Entrada:

```
env_info = {
    'current_turn': int,        # Número de turno actual
    'history': {
        'actions': [int, ...],   # Acciones pasadas (0 para brazo 1, 1
para brazo 2)
        'rewards': [float, ...]  # Recompensas pasadas
    }
}
```

Salida:

```
action = 0 or 1  # 0 para el brazo 1, 1 para el brazo 2
```

5. Métricas de Rendimiento

El entorno evalúa el rendimiento de los agentes usando varias métricas clave:

5.1 Recompensa Promedio

Esta es la recompensa media obtenida por turno, calculada como:

$\text{Recompensa Promedio} = \frac{1}{T} \sum_{t=1}^{T} r_t$

PROF

Esta métrica mide directamente qué tan bien el agente está maximizando su función objetivo. Valores más altos indican un mejor rendimiento.

5.2 Porcentaje de Acciones Óptimas

Esta métrica mide el porcentaje de veces que el agente seleccionó el brazo con la mayor probabilidad de recompensa:

$\text{Acciones Óptimas (%)} = \frac{100}{T} \sum_{t=1}^{T} \mathbf{1}{a_t = \arg\max_i p_i(t)}$

Donde $\mathbf{1}$ es la función indicadora que vale 1 cuando la condición es verdadera y 0 en caso contrario.

Esta métrica muestra con qué frecuencia el agente elige el mejor brazo, independientemente de la recompensa real recibida. Valores más altos indican una mejor selección de brazos.

5.3 Arrepentimiento (Regret)

El arrepentimiento mide la diferencia entre la recompensa esperada de elegir siempre el brazo óptimo y la recompensa esperada de las elecciones del agente:

$\text{Regret} = \sum_{t=1}^{T} \max_i p_i(t) - \sum_{t=1}^{T} p_{a_t+1}(t)$

Valores más bajos de arrepentimiento indican un mejor rendimiento.

5.4 Distribución de Recompensas

El entorno visualiza la distribución de recompensas en diferentes entornos usando diagramas de caja (boxplots) y diagramas de violín (violin plots). Estas visualizaciones ayudan a entender:

- La mediana del rendimiento
- La variabilidad en el rendimiento
- La presencia de valores atípicos
- La forma general de la distribución de recompensas

6. Pautas de Estrategia

6.1 Enfoques Generales

Aquí hay algunos enfoques generales a considerar para la implementación de tus agentes:

1. **Selección Aleatoria**: Elegir brazos aleatoriamente (enfoque de referencia).
2. **Greedy (Codicioso)**: Elegir siempre el brazo con la recompensa estimada más alta.
3. **ε-Greedy**: Casi siempre elegir el mejor brazo, pero explorar ocasionalmente.
4. **UCB (Upper Confidence Bound)**: Elegir brazos basados en estimaciones optimistas de su valor.
5. **Thompson Sampling**: Elegir brazos basados en emparejar probabilidades con distribuciones a posteriori.
6. **Enfoques Bayesianos**: Mantener distribuciones de probabilidad sobre los valores de los brazos.

6.2 Consideraciones Específicas del Entorno

PROF

Banda Fija

- Enfocarse en identificar rápidamente el mejor brazo.
- La exploración se vuelve menos valiosa conforme avanza el juego.
- Con T conocido, se puede planificar un programa decreciente de exploración.

Banda Periódica

- Detectar la estructura periódica (k=10).
- Restablecer estimaciones al comienzo de cada período.
- Asignar más exploración al inicio de cada período.

Banda Dinámica

- Las observaciones recientes valen más que las antiguas.
- Considerar el uso de una ventana deslizante de observaciones.

- Podría necesitar alta capacidad de respuesta a los cambios.

Banda Totalmente Aleatoria

- Estar alerta a cambios repentinos en los patrones de recompensa.
- Equilibrar la persistencia (usar historial) con la adaptabilidad.
- Considerar métodos de detección de cambios.

6.3 Consideraciones Específicas de la Información

Agente de Información Completa

- Aprovechar el valor conocido p1.
- Enfocarse en estimar p2 con eficiencia.
- Ajustar la estrategia dinámicamente con base en los valores relativos.

Agente de Información Parcial

- Similar a información completa, pero más limitado.
- Podría requerir más exploración en ciertos entornos.

Agente de Solo Recompensa

- Debe estimar las probabilidades de ambos brazos.
- Necesita lidiar con el horizonte de tiempo desconocido.
- Considerar estrategias adaptativas en el tiempo.

7. Conclusión

El problema de Multi-Bandas ofrece un marco fundamental para estudiar la toma de decisiones secuenciales bajo incertidumbre. Los entornos y escenarios de información en este playground brindan un conjunto rico de desafíos que resaltan diferentes aspectos del dilema exploración-explotación.

PROF

Al implementar agentes para estos escenarios, obtendrás experiencia práctica con conceptos clave en aprendizaje por refuerzo y teoría de la decisión, y desarrollarás intuición para equilibrar la recolección de información con la maximización de recompensas en diversos contextos.

Mientras trabajas en tus implementaciones, considera cómo se extenderían tus estrategias a:

- Bandas con más de dos brazos.
- Espacios de acción continuos.
- Distribuciones de recompensa no estacionarias con diferentes patrones.
- Bandas contextuales donde se dispone de información adicional.